How Many Cases Do You Need? Assessing and Predicting Case-Base Coverage

David Leake and Mark Wilson

School of Informatics and Computing
Indiana University, Bloomington, IN, USA
leake@cs.indiana.edu, mw54@cs.indiana.edu

Abstract. Case acquisition is the primary learning method for case-based reasoning (CBR), and the ability of a CBR system's case-base to cover the problems it encounters is a crucial factor in its performance. Consequently, the ability to assess the current level of case-base coverage and to predict the incremental benefit of adding cases could play an important role in guiding the case acquisition process. This paper demonstrates that such tasks require different strategies from those of existing competence models, whose aim is to guide selection of competent cases from a known pool of cases. This paper presents initial steps on developing methods for predicting how unseen future cases will affect a system's case-base. It begins by discussing case coverage as defined in prior research, especially in methods based on the representativeness hypothesis. It then compares alternative methods for assessing case-base coverage, including a new Monte-Carlo method for prediction early in case-base growth. It evaluates the performance of these approaches for three tasks: estimating competence, predicting the incremental benefit of acquiring new cases, and predicting the total number of cases required to achieve maximal coverage.

1 Introduction

The case library is a fundamental knowledge container for case-base reasoning (CBR) systems. CBR system development often includes a case acquisition process to capture set of "seed cases," as an initial case-base, after which additional cases are gained during problem-solving. Currently, little quantitative guidance is available to help system-builders to predict the likely benefit of acquiring another seed case and the number of cases which will be required to maximize coverage. The ability to predict the incremental benefit of case acquisition could help to determine whether the effort to acquire new cases is worthwhile; predictions of the case-base size required to achieve a desired performance level—and of whether it is practical to achieve that performance level solely by adding cases—could both help to determine the suitability of CBR for a given task and help system designers decide whether to focus effort on case acquisition or on improving other knowledge containers, such as case adaptation knowledge, to reduce the number of cases that will be needed.

Assessing and predicting the effects of case acquisition is closely connected to estimating *case-base competence*, the ability of a system's case-base to support the solution of potential target problems. Methods to predict case-base competence effects if unseen cases are added to the case-base can in turn be used to estimate incremental competence gains from adding a future case to the case-base. If methods can be developed to extrapolate these predictions to estimate the competence effects of adding larger numbers of cases, those estimates may be used for addressing questions such as the maximum competence the system is likely to achieve.

Case-base competence has received substantial attention in CBR research, from the perspective of guiding generation of compact and competent case-bases. This work has centered on choosing which cases in an existing set of cases to delete (for competence-based deletion, e.g., [4, 6]) or to add (for competence-based addition, e.g., [7]). This paper begins with background on previous treatments of competence, based on the representativeness assumption that the existing case-base can be used as a proxy for future problems [6]. Such treatments have proven effective for their intended purpose of guiding competence-based deletion from a known case-base with satisfactory coverage. However, representativeness is not assured for the partial case-bases arising during early phases of case acquisition.

To estimate coverage characteristics, the paper presents a set of empirical methods for future coverage prediction. These include a new approach which uses Monte Carlo integration—assessing coverage for a random sampling of the problem space—to predict coverage without requiring the representativeness assumption. This approach does not require knowledge of the correct solutions for the sampled problems, but can be adjusted to reflect additional knowledge about expected problem distributions and weighted to reflect additional cost/benefit information.

The paper presents an evaluation of its competence estimation methods as bases for three tasks: (1) predicting the incremental benefit of case acquisition, (2) estimating a tight upper bound on the competence a system will achieve, based on the competence effects of adding initial cases, and (3) predicting the number of cases required for the case base to approximate maximal competence. The paper evaluates performance for a range of case-bases. The results support that representativeness-based methods for competence assessment may not be well-suited to early competence estimation, but that both the leave-one-out and Monte Carlo methods can provide useful information, and that the Monte Carlo method has advantages over the leave-one out method for this task.

2 Competence and Representativeness

Smyth and McKenna's seminal work on representativeness-based coverage [6] defines competence as "the range of target problems that a given system can solve." If "competence" is considered to reflect problem-solving accuracy over the entire problem space P, it may be defined as the fraction of the problems

$p_i \in P$ which the reasoner will solve correctly, i.e., for which the case or cases retrieved from the case-base will be adapted to produce a correct solution:

$$Competence(CB) = \frac{\sum_{p_i \in P} Correct(Adapt(Retrieve(CB,\ p_i),\ p_i))}{|P|}$$

A challenge for determining competence for a real problem distribution is that the actual set of problems to be encountered in the future may be impossible to know *a priori*. In the context of determining the contribution of particular cases to the competence of a given case base, Smyth and McKenna address the problem of unavailable target cases by basing competence calculations on the *representativeness assumption*: "The case base is a representative sample of the target problem space" [6]. As Smyth and McKenna observe, in the scenario they consider it is reasonable to expect the representativeness hypothesis to hold: If the case-base were not representative of future problems, CBR would be inappropriate for the task. Accordingly, their proposed competence metric explicitly excludes mention of target problems, instead considering only how the system's cases contribute to solving other cases in its existing case-base. The coverage of a single case c with respect to a case-base C is defined as the set of cases for whose problems it would be retrieved and to which it can be successfully adapted [5]:

$$coverage(c \in C) = \{c' \in C : c' \in RetrievalSpace(c) \cap AdaptationSpace(c)\}$$

The representativeness-based approach has been used as a basis for estimating both global competence [5] and relative coverage [6], a criterion for determining which cases are most important to retain in the case-base. Representativeness approaches have been shown to be effective for guiding competence-preserving case deletion from a set of cases with satisfactory competence. However, during initial case acquisition, before satisfactory coverage is achieved, there is no guarantee that the cases seen will be representative.

Some prior work has studied how to identify additional cases needed for finite case-bases [3]. However, there has been little attention to the problem of predicting the number of additional cases which a CBR system may need to achieve maximal competence. It might appear that prior approaches for assessing case-base competence could also be used to predict the number of cases needed. However, Massie, Craw, and Wiratunga [2] have shown that metrics developed to assess competence are not necessarily good indicators of the accuracy achievable with a given set of cases.

3 Empirical Competence Estimation Strategies

We consider four empirical approaches for estimating competence: leave-one-out testing using a limited number of test cases, leave-one-out testing using all cases currently in the case-base, competence group estimation, and a Monte Carlo method.

3.1 Leave one Out Testing

Leave-one-out testing is a simple approach for estimating accuracy. For tasks for which solvability is Boolean—either a case is solved or it is not—competence depends solely on the percent of target problems solved correctly. For other types of problems, leave-one-out testing can be applied in conjunction with other types of criteria for estimating competence, such as determining the percent of target problems whose solution is within a given threshold of the correct solution, or even simply considering average solution accuracy. Other criteria could use different weightings for different problems (e.g., based on the risks associated with particular types of errors).

We will consider two variants on leave-one-out testing. The first uses all the cases in the current case-base. The second increases efficiency by conducting testing using a smaller subset of the case-base.

3.2 Competence group estimation

The representativeness-based competence metric we consider follows Smyth and McKenna's coverage metric [6], based on the notion of competence groups. Each case c has an associated *coverage* within the case-base, consisting of those cases which are retrieved for c by the reasoner's retrieval component and which can be adapted to solve c by the reasoner's adaptation component. A *competence group* is defined as a set of cases such that all cases share coverage with some other case in the group, and no case outside the group shares coverage with any case in the group. The *density* of a case c within a group G is defined as:

$$\text{Density}(c, G) = \frac{\sum_{c' \in G - \{c\}} \text{Similarity}(c, c')}{|G| - 1}$$

The *group density* is defined as the sum of the individual member cases' densities, divided by the cardinality of the group. *Group coverage* is taken as:

$$\text{GroupCoverage}(G) = 1 + (|G| \cdot (1 - \text{GroupDensity}(G)))$$

Case-base coverage is taken as the sum of group coverages over all competence groups in the case-base.

3.3 Monte Carlo Coverage Estimation

Leave-one-out testing and representativeness-based approaches use existing cases in the case-base as proxies for future problems. When the available cases may not be representative, as when few cases have been acquired, this assumption is less appropriate. To predict coverage of cases which may not be reflected by the existing case-base, we propose a Monte Carlo technique. This Monte Carlo method uniformly samples the problem space, and tests whether the sampled points are expected to be solvable. This process does not require actually generating solutions for the sampled points, provided that a criterion exists for determining

1. Generate problem set with desired distribution
2. Initialize the total cost to 0
3. For each problem p:
   3a. Find the closest case to p in the case-base
   3b. If that case does not cover p, add the cost of not covering p to the total cost.

Table 1: General Monte Carlo sampling algorithm

Apply General Monte Carlo sampling algorithm with:
   Uniform problem distribution for n samples
   Cost = 0 if problem covered by nearest case; else 1.
Return (Monte Carlo result)/n

Table 2: Monte Carlo coverage estimation algorithm for our experiments

solvability. For example, a sampled case could be considered solvable if it were sufficiently similar to an existing case in the case-base, based on the system similarity metric. Because this method does not require access to any cases beyond those already in the case-base, the number of points it can test is limited only by available processing time. This is contrast to the leave one out approaches, which are limited by the number of cases in the case-base.

The previously-described Monte Carlo process can be enriched in two ways to better reflect pragmatic constraints. We describe these for generality, but leave their exploration for future research:

- **Biased sampling:** When the problems encountered are not uniformly distributed, and if information about the distribution is available, the Monte Carlo sample selection process can be biased to reflect that distribution. The most informative results about the system's ability to cover problems in practice would follow from sampling frequently from regions of the problem space in which future problems are likely to occur and less frequently in problem areas that are unlikely to arise.
- **Problem-specific costs:** Rather than simply considering points as "covered" or not, a cost function could be used to reflect factors such as finer-grained accuracy loss or the costs of failure to cover particular cases, based on knowledge of the importance of those cases, as illustrated in Table 1.

In the following, we apply the general Monte Carlo algorithm using a problem generator which randomly selects problems with a uniform distribution throughout the problem space. Cases which fall within the coverage of an existing case, as determined by a similarity threshold, are recorded as solved. The percent solved is then be used to approximate the coverage of the case-base. Table 2 sketches our general Monte Carlo Coverage algorithm and its application here.

4 Extrapolating from Competence Graphs to Estimate Needed Cases

As illustrated in the following experiments, the observed coverage growth for our sample case-bases followed a standard pattern, reaching an asymptotic value. If the details of this standard pattern can be predicted for a given case-base, such predictions can be applied to in turn predict the number of cases needed to approximate this maximal performance level.

We hypothesized that the general form of competence as a function of case-base size (x) can be approximated by the following function:

$$f(x) = c \cdot (1 - (x + b)^{-p})$$

The shape of the function is displayed by the fitted curves shown in Fig. 2 (all curves shown except for the Empirical Accuracy graph, which represents raw data points).

This function captures the "elbow" or corner point of diminishing returns that is typical of these experiments. Early in case acquisition, insufficient data will be available to make reasonable long-term predictions. However, we hypothesized that prediction algorithms can detect when the "elbow" of the curve is reached, and at that point prediction can begin.

5 Experiments

5.1 Overview and Design

We conducted experiments to compare the performance of representativeness, leave-one-out, and Monte-Carlo integration as a basis for the following tasks:

1. Estimation of system competence
2. Prediction of the marginal competence benefit from acquiring a new case
3. Prediction of the number of cases needed to for the system to achieve accuracy within ϵ of its maximum accuracy level

Our experiments use four classification data sets, drawn from the UCI Machine Learning Repository [1]: Ad (classification of Internet images as ads), MiniBOONE (classification of particles), Adult (classification of income level), and Car (classification of car models as acceptable to consumers). For each data set, the same naive similarity metric was used, Euclidean distance on problem features normalized by their ranges. Data sets with categorical features were given simple hand-designed numeric distances between categories for those features.

The experiments used 10-fold cross-validation, with each data set split randomly into $f = 10$ folds. Tunable experiment parameters included the range of case-base size to test and the size increments to use, the number of points for the Monte Carlo procedure to sample (for our experiments, 50 points were sampled), and the number of nearest cases to use when generating a CBR solution for test problems (3-NN for our experiments).

Estimation of Competence The experiments test competence estimation for a variety of case-base sizes. For each case-base size, the cases for the case-base are drawn sequentially from an initial random ordering of the current fold. At each case-base size step, five values are computed:

1. Empirical accuracy: The percentage of problems from the $(f-1)$ test folds which are solved correctly by the case-base
2. A leave-one-out estimate of the case-base's accuracy, using all cases in the current case-base
3. A leave-one-out estimate of the case-base's accuracy, limited to the same number of samples as the Monte-Carlo estimate
4. The representativeness coverage value
5. The Monte Carlo estimate of the case-base's coverage

The comparative results of 3 and 4 are interesting in that they enable comparing the effectiveness of leave-one-out and Monte Carlo methods when each has access to the same amount of information.

Prediction of marginal coverage benefit of next case addition As the basis for prediction of the marginal coverage benefit, the system attempts to fit the values produced by each estimation technique to the curve described in Section 4. To fit the data to the curve, each set of values estimating the competence of the case-base is linearized according to the inverse of the previously stated curve, *i.e.*, by:

$$f^{-1}(y) = (1 - \frac{y}{c})^{-1/p} - b$$

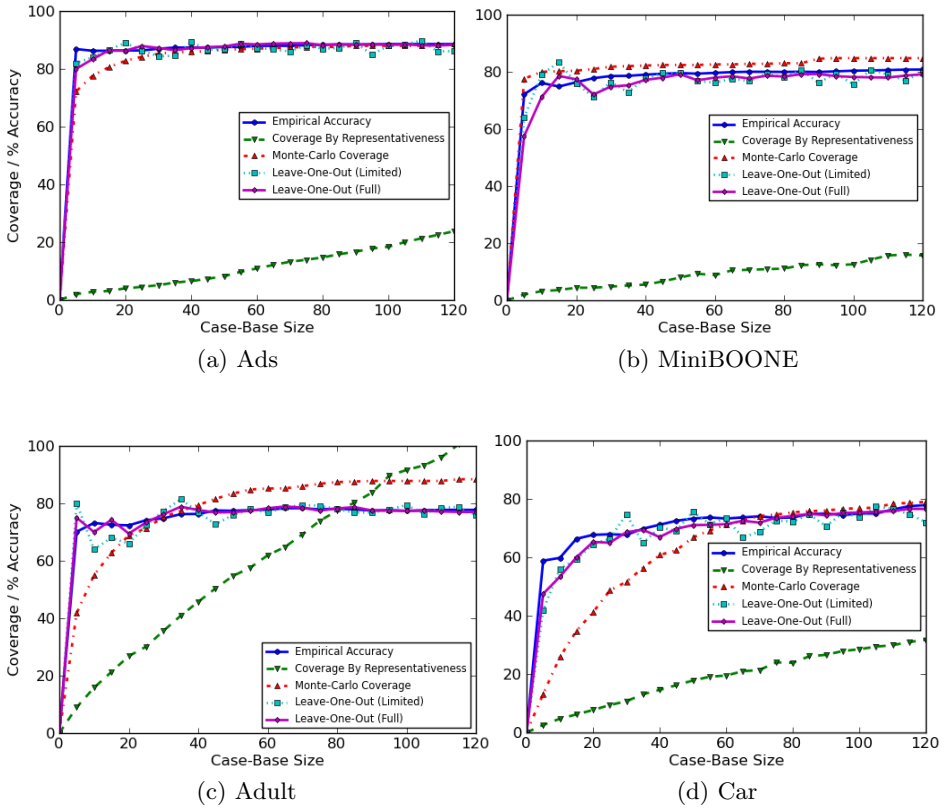
Where $\mathbf{y}$ are the estimate values. A least-squares fitting is used to fit the parameters c, b, and p to the known case-base sizes $\mathbf{x}$.

The curve fitted to the first s points is used to predict the gain in accuracy that will result by expanding the case-base to size $s+i$.

Prediction of the Number of Cases Required for Maximal Coverage The fitted curve can also be used to predict the number of cases that will be required for the reasoner to reach within ϵ of a target accuracy value a. In our experiments, we use the curve fitted to the estimate values up to s to predict the case-base size at which the reasoner will reach an accuracy of at least $a - \epsilon$.

For our experiments, we set a to the empirical accuracy value obtained with the largest case-base size sampled and set ϵ to 5% of a. In practice, a developer could select any desired target accuracy value less than the maximum.

Predictions are only generated after the "elbow" of the curve has been crossed. The number of cases needed to reach a is re-predicted after each acquisition step, as with the marginal-benefit task, allowing the estimation methods to refine the prediction of the needed number of cases after every acquisition. We expect the accuracy of all prediction methods to improve (on average) with each case acquired.



(a) Ads (b) MiniBOONE

(c) Adult (d) Car

Fig. 1: Estimates of coverage, accuracy, and empirical accuracy

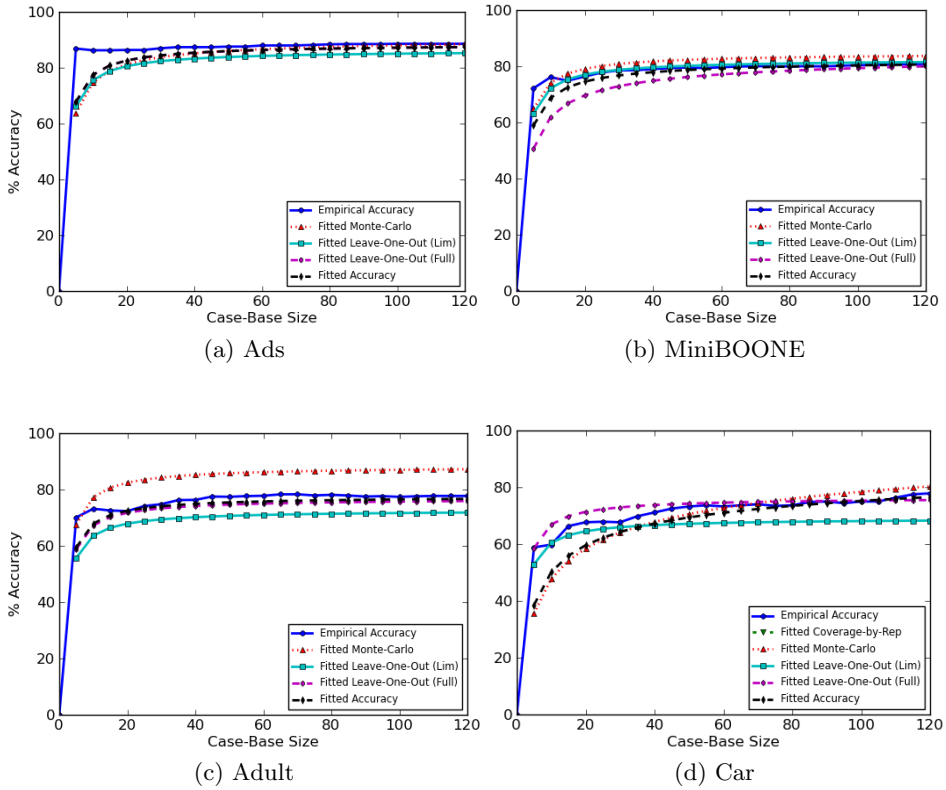
By comparing predicted values to the empirical results – i.e., the gain in empirical accuracy by expanding the case-base to size $s+i$ and the size at which the case-base's empirical accuracy crossed $a-\epsilon$, we determined the accuracy of these predictions for each estimation technique at every case-base size evaluated.

5.2 Results

Estimation of Competence Fig. 1 shows the estimation results and empirical accuracy. We note that the representativeness function is not intended to produce a result in percent accuracy, so it is only meaningful to compare its shape to the shapes of the curves for the other methods. We observe that for both Ads and MiniBOONE, maximal accuracy is approached with a small number of cases. Leave-one-out and Monte Carlo methods both track actual accuracy closely after an initial lag and level off quickly. Results with Adult and Car show

more differentiation between the methods, with full leave-one-out providing the best performance, followed by limited leave-one-out and then Monte Carlo.

The general behavior of the representativeness coverage estimates contrasts with that of other methods, producing linear or nearly linear growth as a function of case-base size. This is observation on representativeness estimates is consistent with results by Massie et al. [2].

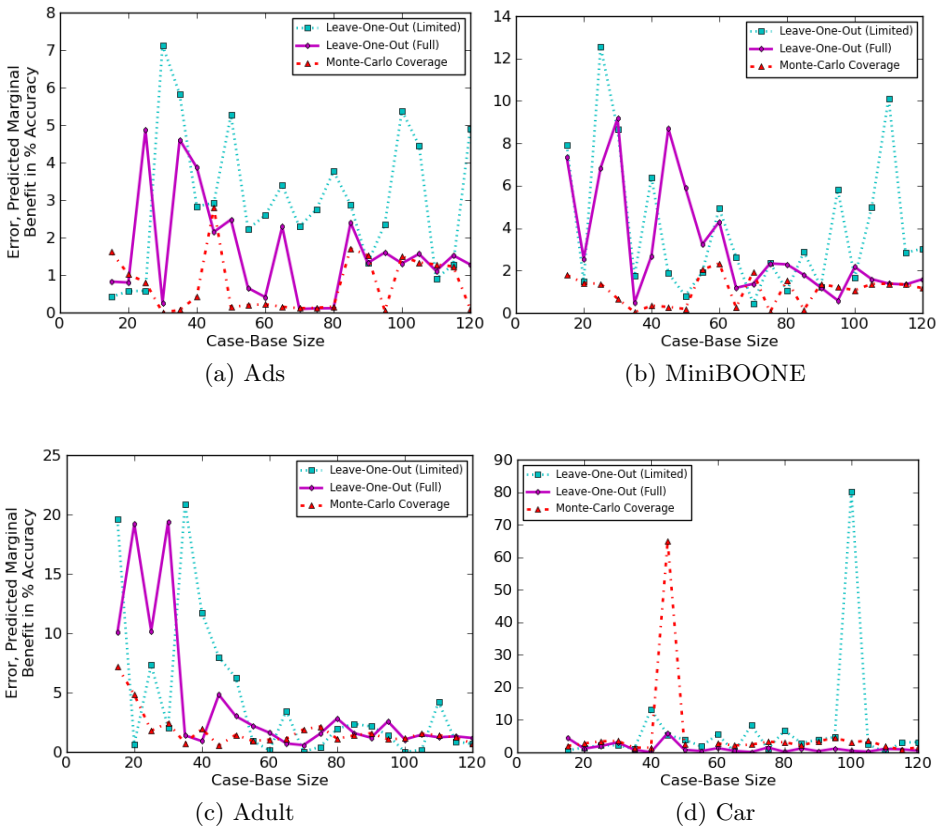


(a) Ads (b) MiniBOONE

(c) Adult (d) Car

Fig. 2: Empirical accuracy and curves fitted to the estimates of accuracy/coverage

Curve Fitting to Accuracy Estimates Because the general form of the representativeness graph does not match the curve of the other methods, we consider only the results for the other methods. Fig. 2 shows the results of fitting the curve to the other methods, which all fit the general pattern, with some variation compared to empirical accuracy.



(a) Ads (b) MiniBOONE

(c) Adult (d) Car

Fig. 3: Error in predicted marginal benefit of case acquisition, by case-base size

Prediction of Marginal Benefit of Acquisition Fig. 3 shows the absolute error in predicted marginal benefit of case acquisition, for each case-base size for which predictions are available. (For some case-base sizes, curve fitting was occasionally unsuccessful for some estimate methods. Missing values in the graph reflect failed curve-fitting, and the estimate technique incurs no error penalty for these missed predictions.) Fig. 4 graphs the means of the available error values for three different regions of case acquisition – early, middle, and late case-base growth. To compute these values, the entire experiment was split evenly into three stages and averages computed for each stage, to illustrate the accuracy of different estimate techniques at each stage. These values illustrate that the Monte-Carlo integration method generally compares favorably with leave-one-out for predicting marginal benefit of new case acquisitions. In the Car data set, the Monte-Carlo technique bests the leave-one-out technique in two out of three stages when the leave-one-out is limited to the same number of samples as

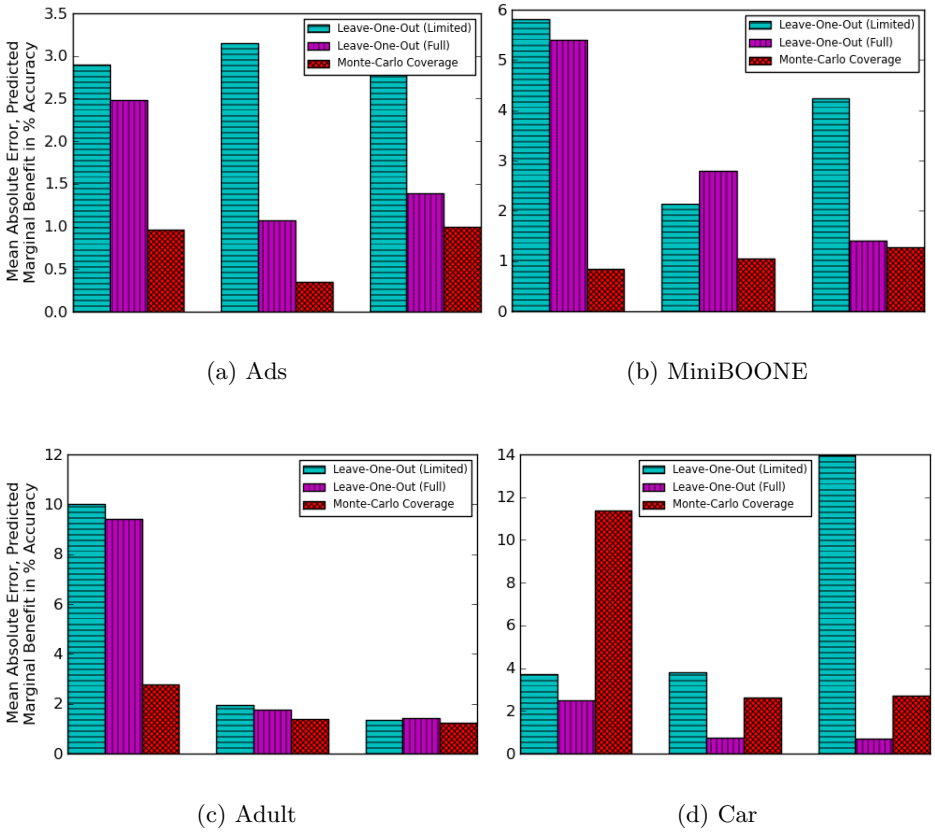
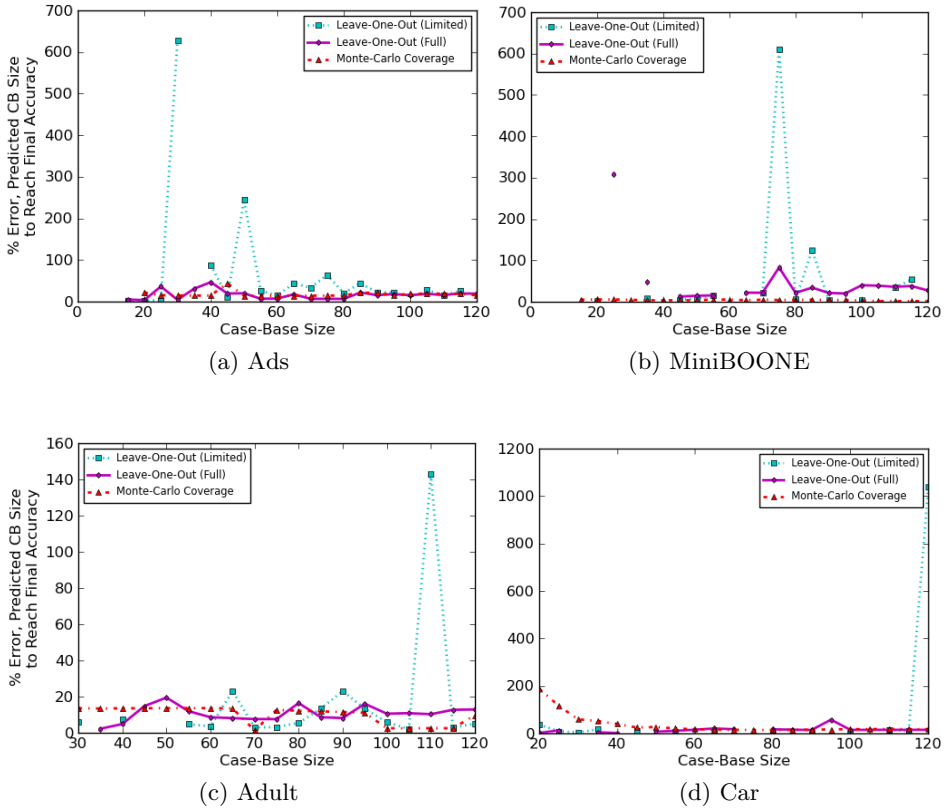


Fig. 4: Mean error in predicted marginal benefit of case acquisition, for early, middle, and late stages of case-base growth

Monte-Carlo, but not when leave-one-out is allowed the full range of the case-base. However, see below for a discussion of the time required to execute each test.

Prediction of Number of Cases Needed to Achieve Maximal Accuracy The absolute error in predicting the case-base size required to reach at least within ϵ of the final experimental accuracy is shown in Fig. 5. These values are presented as percentages of the final case-base size in the experiment. The mean absolute error in these predictions is shown for each data set in Fig. 6. The error in the Monte-Carlo technique is higher here, but it is often possible to produce a prediction with the Monte-Carlo method when such a prediction is impossible with the leave-one-out techniques because a curve could not be fitted.



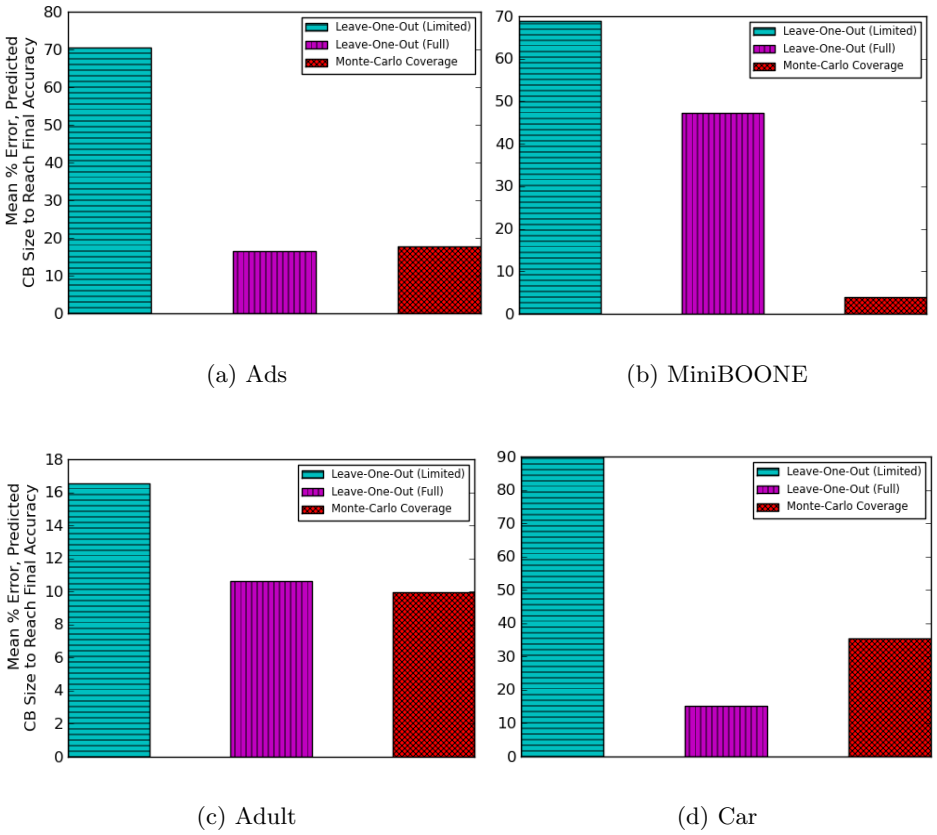
(a) Ads (b) MiniBOONE

(c) Adult (d) Car

Fig. 5: Error in predicted case-base size to reach within ϵ of final experimental accuracy, by case-base size

After 120 cases, the respective errors for limited leave-one-out, full leave-one-out, and Monte Carlo, for Ads are no prediction possible, 19%, and 15%; for MiniBOONE are no prediction possible, 27%, and 2%; for Adult are 5%, 13%, and 10%; and for Car are 1038%, 15%, and 18%.

Note on Computation Time The time elapsed to compute the estimates with each technique is shown in Fig. 7. The Monte-Carlo coverage method required less time than the representativeness coverage technique or the leave-one-out estimate using the full case-base (although leave-one-out can be faster for very small case-bases, its time grows more quickly and rapidly overtakes the Monte-Carlo technique). When leave-one-out testing is limited to the same number of samples as the Monte-Carlo technique, their elapsed time is comparable; however, as shown by the previous results, the accuracy of the leave-one-out technique is generally compromised by doing so.



(a) Ads

(b) MiniBOONE

(c) Adult

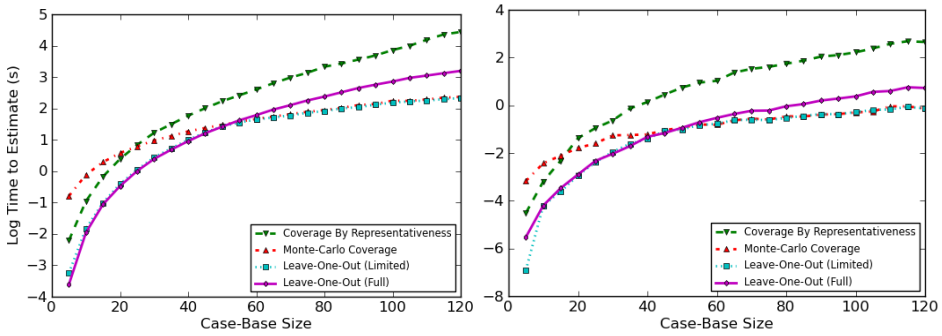
(d) Car

Fig. 6: Mean error in predicted case-base size to reach within ϵ final experimental accuracy

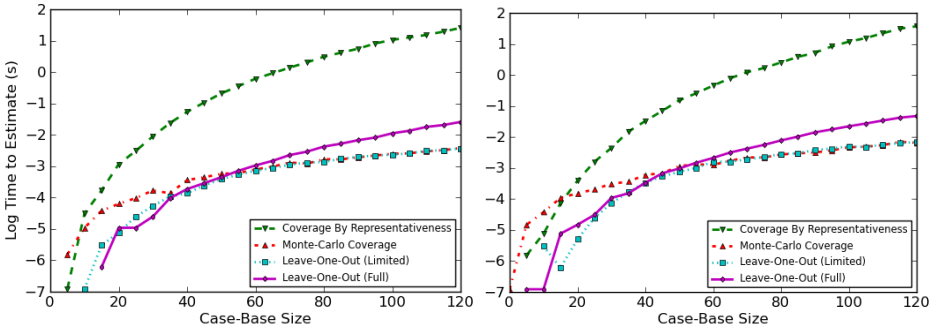
6 Future Work

The previous sections introduce the problem of predicting case-base coverage, illustrate some central points, and present experiments testing initial methods. A number of questions remain. One is how best to handle problem streams with non-uniform distributions, if those distributions are not known *a priori*. Another interesting future area is how to develop automated methods for selecting values such as similarity thresholds for deciding whether to treat a case as covered.

The ability to predict the benefits of case acquisition also raises questions for the tradeoff between increased case adaptation knowledge and increased case knowledge and how to provide guidance for CBR system developers deciding how to divide their effort between augmenting these two knowledge containers.



(a) Ads (b) MiniBOONE



(c) Adult (d) Car

Fig. 7: Log time (in seconds) to compute accuracy and competence estimates

7 Conclusion

As the acquisition of seed cases is an important part of the development of CBR systems, the ability to predict the benefit of such acquisitions could play a valuable role in guiding case acquisition decisions. Likewise, knowledge of the benefit trends for case acquisition can aid in predicting the number of cases which will be needed to achieve a desired level of accuracy and in predicting limits on the accuracy attainable, aiding predictions of the practicality and effort required to build a CBR system.

This paper explores methods for predicting coverage growth, including a Monte Carlo simulation method to enable predictions early in the case acquisition process, and presents tests illustrating the methods potential. This work provides a first step towards answering the question of how to predict the number of cases it will be necessary to acquire for a CBR system.

Acknowledgments

This work is based on work supported by the National Science Foundation under Grant No. OCI-0721674 and by a grant from the Data to Insight Center of Indiana University.

References

1. Blake, C., Merz, C.: UCI repository of machine learning databases (1998), http://www.ics.uci.edu/∼mlearn/MLRepository.html
2. Massie, S., Craw, S., Wiratunga, N.: Complexity-guided case discovery for case based reasoning. In: AAAI'05: Proceedings of the 20th national conference on Artificial intelligence. pp. 216–221. AAAI Press (2005)
3. McSherry, D.: Automating case selection in the construction of a case library. Knowledge-Based Systems 13(2-3), 133–140 (2000)
4. Smyth, B., Keane, M.: Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. pp. 377–382. Morgan Kaufmann, San Mateo (August 1995)
5. Smyth, B., McKenna, E.: Modelling the competence of case-bases. In: Cunningham, P., Smyth, B., Keane, M. (eds.) Proceedings of the Fourth European Workshop on Case-Based Reasoning. pp. 208–220. Springer Verlag, Berlin (1998)
6. Smyth, B., McKenna, E.: Building compact competent case-bases. In: Proceedings of the Third International Conference on Case-Based Reasoning. pp. 329–342. Springer Verlag, Berlin (1999)
7. Zhu, J., Yang, Q.: Remembering to add: Competence-preserving case-addition policies for case base maintenance. In: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence. pp. 234–241. Morgan Kaufmann (1999)